Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5И-21М
Ван Чжэн

Москва — 2024 г.

1. Цель лабораторной работы

Изучить различные методы визуализации данных [1].

2. Задание

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

- Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

- Сформировать отчет и разместить его в своем репозитории на github.

3. Ход выполнения работы

3.1.Описание набора данных

В этой тетради я буду использовать графики для визуализации взаимосвязи между переменными в наборе данных "Качество воздуха из Нью-Йорка".

О наборе данных:

Набор данных содержит информацию о данных наблюдения за качеством воздуха в Нью-Йорке.

Загрязнение воздуха является одной из наиболее важных экологических угроз для городского населения, и, хотя воздействию подвергаются все люди, выбросы загрязняющих веществ, уровни воздействия и уязвимость населения различаются в зависимости от района.

3.2. Основные характеристики набора данных

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv('/content/Air_Quality.csv')
df
```

	Unique ID	Indicator ID	Name	Measure	Measure Info	Geo Type Name	Geo Join ID	Geo Place Name	Time Period	Start_Date	Data Value	Message
0	179772	640	Boiler Emissions- Total SO2 Emissions	Number per km2	number	UHF42	409.0	Southeast Queens	2015	01/01/2015	0.3	NaN
1	179785	640	Boiler Emissions- Total SO2 Emissions	Number per km2	number	UHF42	209.0	Bensonhurst - Bay Ridge	2015	01/01/2015	1.2	NaN
2	178540	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	209.0	Bensonhurst - Bay Ridge	Annual Average 2012	12/01/2011	8.6	NaN
3	178561	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409.0	Southeast Queens	Annual Average 2012	12/01/2011	8.0	NaN
4	823217	365	Fine particles (PM 2.5)	Mean	mcg/m3	UHF42	409.0	Southeast Queens	Summer 2022	06/01/2022	6.1	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
18020	816914	643	Annual vehicle miles traveled	Million miles	per square mile	CD	503.0	Tottenville and Great Kills (CD3)	2019	01/01/2019	12.9	NaN
18021	816913	643	Annual vehicle miles traveled	Million miles	per square mile	CD	503.0	Tottenville and Great Kills (CD3)	2010	01/01/2010	14.7	NaN
18022	816872	643	Annual vehicle miles traveled	Million miles	per square mile	UHF42	208.0	Canarsie - Flatlands	2010	01/01/2010	43.4	NaN
18023	816832	643	Annual vehicle miles traveled	Million miles	per square mile	UHF42	407.0	Southwest Queens	2010	01/01/2010	65.8	NaN
18024	151658	643	Annual vehicle miles traveled	Million miles	per square mile	UHF42	408.0	Jamaica	2005	01/01/2005	41.0	NaN

18025 rows × 12 columns

```
data.dtypes
```

```
Unique ID          int64
Indicator ID       int64
Name              object
Measure           object
Measure Info      object
Geo Type Name     object
Geo Join ID      float64
Geo Place Name    object
Time Period       object
Start_Date        object
Data Value       float64
Message          float64
dtype: object
```

```
plt.figure(figsize=(30, 40))
average_data_value = df.groupby('Geo Place Name')['Data Value'].mean().reset_index()
sns.barplot(x='Data Value', y='Geo Place Name', data=average_data_value)
plt.title('Average Data Value by Geo Place Name')
plt.xlabel('Average Data Value')
plt.ylabel('Geo Place Name')
plt.show()
```

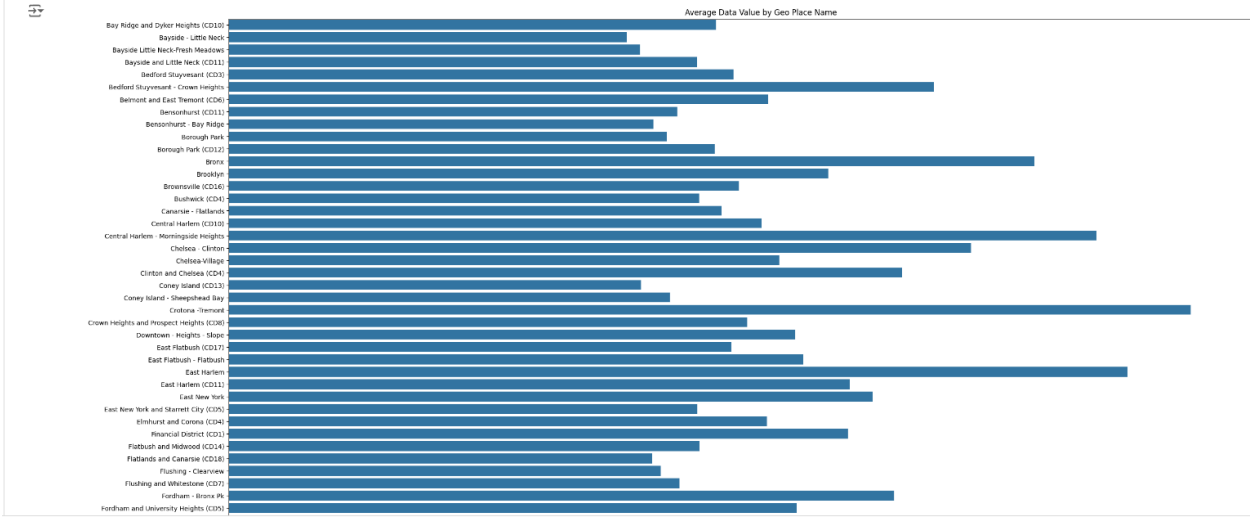

Рис 1. -Среднее значение данных по географическим названиям

```
# Convert Start_Date to datetime
df['Start_Date'] = pd.to_datetime(df['Start_Date'])

# Create scatter plot
plt.figure(figsize=(12,6))
plt.scatter(df['Start_Date'], df['Data Value'])
plt.title('Air Pollutant Levels Over Time')
plt.xlabel('Date')
plt.ylabel('Pollutant Level')
plt.tight_layout()
plt.show()
```

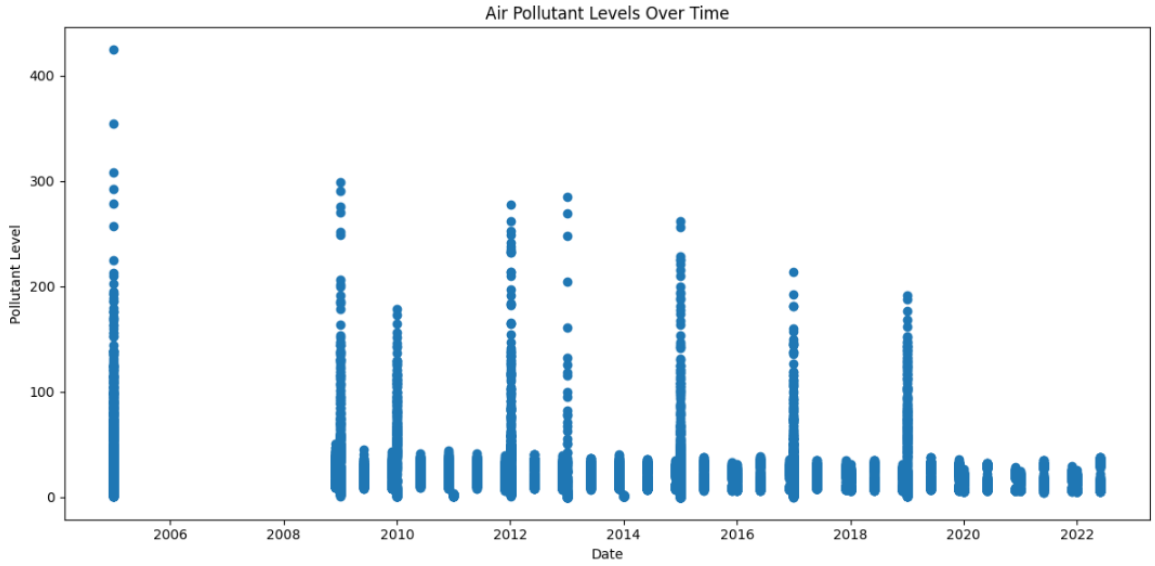

Рис 2. - Уровни загрязнителей воздуха с течением времени

```
numeric_df = df.select_dtypes(include=[np.number])

# Calculate correlation matrix
corr_matrix = numeric_df.corr()

# Create heatmap
plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='YlOrRd')
plt.title('Correlation of Air Pollutants')
plt.tight_layout()
plt.show()
```

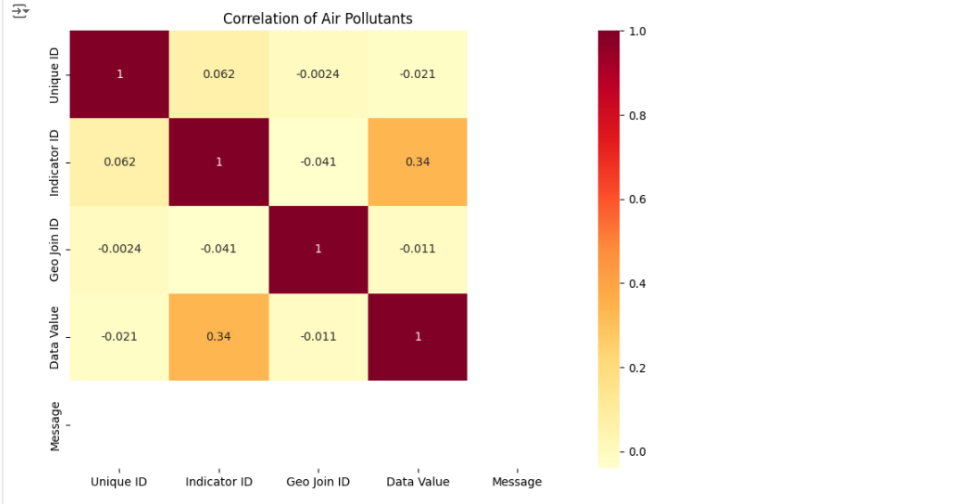

Рис 3. - Корреляция загрязнителей воздуха

```
[18] plt.figure(figsize=(14, 7))
     sns.boxenplot(x='Measure Info', y='Data Value', data=df)
     plt.title('Boxen Plot of Data Value by Measure Info')
     plt.xlabel('Measure Info')
     plt.ylabel('Data Value')
     plt.xticks(rotation=45)
     plt.show()
```

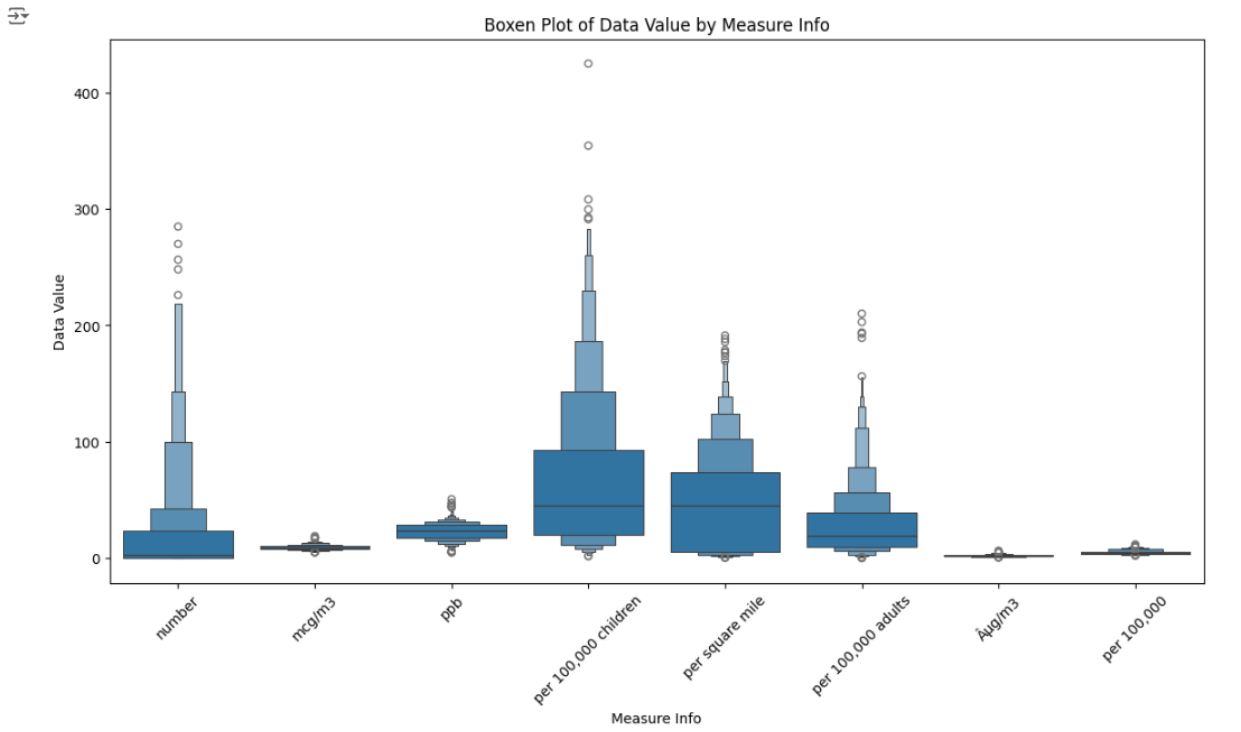

Рис 4. - График Боксена значения данных по мере информации

```python
# @title Measure Info vs Geo Type Name
from matplotlib import pyplot as plt
import seaborn as sns
import pandas as pd
plt.subplots(figsize=(8, 8))
df_2dhist = pd.DataFrame({
       x_label: grp['Geo Type Name'].value_counts()
       for x_label, grp in df.groupby('Measure Info')
})
sns.heatmap(df_2dhist, cmap='viridis')
plt.xlabel('Measure Info')
_ = plt.ylabel('Geo Type Name')
```

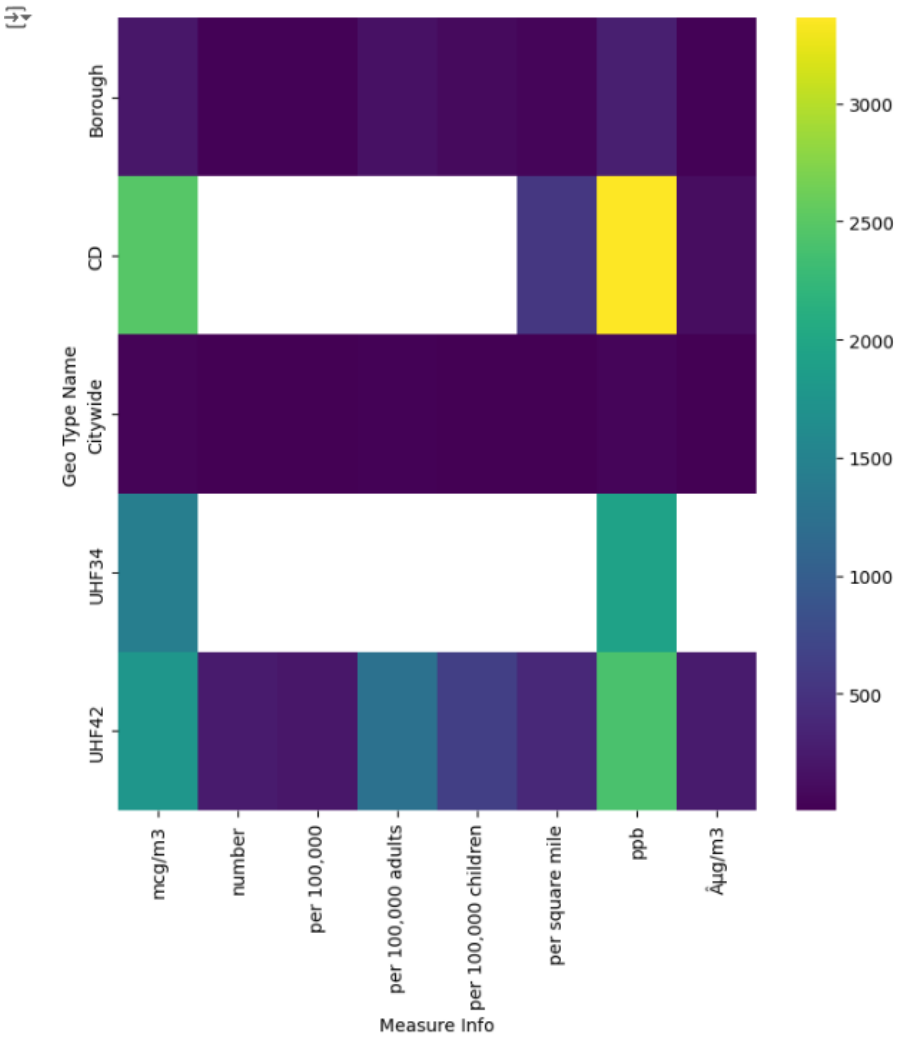

Рис 5. - Измерение информации по названию типа GEO

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets